영한 번역기의 상용화를 위한 도메인 특화 방법의 진화

최승권○, 이기영, 노윤형, 권오욱, 김영길
한국전자통신연구원 언어처리연구팀
{choisk, leeky, yhroh, ohwoog, kimyk}@etri.re.kr

Evolution of Customization Method for Commercialization of an English-Korean MT System[1)]

Sung-Kwon Choi○, Ki-Young Lee, Yoon-Hyung Roh, Oh-Woog Kwon, and Young-Gil Kim
Natural Language Processing Team, ETRI

요 약

본 논문은 한국전자통신연구원 언어처리연구팀에서 2004년까지 개발하였던 웹문서 자동번역 시스템을 2006년부터 매년 도메인별로 상용화에 성공한 사례를 기술한 것이다. 상용화가 가능하였던 주요 요인 중 하나인 도메인 특화 방법을 소개하며 이 도메인 특화 방법이 시기별로 개선되어 진화되는 모습을 기술한다. 즉 2004년의 웹문서 영한 자동번역기를 2006년에 특허문서 영한 자동번역기로 특화할 때 사용한 도메인 특화 방법이 '초기 도메인 특화 방법'이라 할 수 있는데, 이 초기의 도메인 특화 방법에 번역지식 및 번역엔진 모듈의 반자동 튜닝 방법과 자동화된 평가 방법을 추가하여 2007년에 '개선된 도메인 특화 방법'을 개발하였다. 이 '개선된 도메인 특화 방법'은 2007년에 특허문서 영한 자동번역기를 기술논문 영한 자동번역기로, 2008년에 기술논문 영한 자동번역기를 IT웹신문 영한 자동번역기로, 2009년에 IT 웹신문 영한 자동번역기를 전자우편 및 기업문서 영한 자동번역기로, 그리고 2010년에 전자우편 영한 자동번역기를 메신저 영한 자동번역기로 구현할 때 사용하였으며 그 효과는 신규 도메인용 영한 번역기를 개발하는 기간을 점차적으로 줄이게 하였으며 구현 프로세스에 일관성을 제공하였다.

주제어: 자동 번역, 기계 번역, 영한 번역, 도메인 특화

1. 서론

현재 세계적으로 자동번역 시스템의 번역 방식은 크게 규칙기반 자동번역 방식과 통계기반 자동번역 방식으로 구분되어 있다. 통계기반 자동번역 방식이 규칙기반 자동번역 방식에 비해 개발 기간을 절약할 수 있고, 특정 언어에 제약 없이 자동번역 시스템을 개발 할 수 있다는 장점이 있다. 하지만 번역률 관점에서 볼 때 동일 어족에서는 번역률이 뛰어나지만 이종 어족에서는 번역률이 규칙기반 방식보다 떨어지는 경향이 있다.

이종 언어 사이에서 통계기반 방식이 규칙기반 방식보다 상용화가 뒤쳐지는 이유로, 우선 자료 희소성을 들 수 있으며, 두 번째는 언어적 차이를 들 수 있다. 통계기반 방식은 기본적으로 대량의 병렬코퍼스를 토대로 번역 정보를 추출하나 병렬코퍼스가 부족할 때에는 번역 정보를 추출하지 못하기 때문에 번역률이 낮게 된다. 따라서 자료 희소성 문제를 해결하고 고품질의 번역률을 내기 위해서는 병렬 코퍼스를 대량으로 구축하여야 하는데 병렬 코퍼스의 구축 비용 문제가 통계기반 방식의 단점 중 하나라고 할 수 있다. 둘째로 이종 언어는 동종 언어보다 언어적 차이가 더욱 크게 나타난다. 예를 들면 영어와 한국어간에는 다른 동종 언어보다 어순, 원거리 의존관계(long-distance dependency)와 같은 언어 현상이 통계 기반 정보 추출을 어렵게 하는 요인이 된다.

이 때문에 영한 자동번역기는 통계기반 자동번역 방식보다는 패턴을 규칙으로 한 패턴기반 자동번역 방식으로 상용화에 성공을 거둘 수 있었다. 패턴기반 영한 자동번역기는 초기에 그 응용 분야를 영어 웹문서로 하여 개발되었으나 개방된 도메인 때문에 번역률이 낮아 일반 사용자의 외면을 받았었다. 그러나 그 후 그 응용분야를 줄임으로써 상용화에 성공하게 되었다. 상용화에 성공한 사례로는 특허문서[1], 과학 기술 논문[2], IT 웹신문[3], 기업문서[4]등이었으며 현재는 전자우편과 Microsoft 메신저를 대상으로 개발되고 있다.

본 논문의 목표는 패턴기반 영한 자동번역 시스템을 상용화에 성공하게 한 주요 요인인 도메인 특화 방법에 대해 소개하며 도메인별로 적응하였던 도메인 특화 방법의 진화하는 모습에 대해 기술하는 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 웹에서 특허로 특화한 초기 도메인 특화 방법을 기술하고자 한다. 3장에서는 초기 도메인 특화 방법의 문제점을 기술하고, 4장에서는 특허 번역기를 기술논문 번역기로 특화하는데 사용한 개선된 도메인 특화 방법을 소개하고자 한다. 그리고 5장에서는 개선된 도메인 특화 방법이 타당하였다는 것을 새로운 도메인인 전자우편 번역기로부터 메신저 번역기에 적용하여 효과를 보이고자 한다.

2. 초기 도메인 특화 방법: 웹 번역기를 특허 번역기로 특화하기

1) 본 논문은 지식경제부의 산업원천기술 개발사업 (2011-S-034-01)의 일환으로 개발된 결과임을 밝힙니다.

한 도메인으로부터 다른 도메인으로 시스템을 적응해 가는 방법이 도메인 특화 방법이다. 다국어 자동번역 시스템인 SYSTRAN을 대상으로 일반 도메인으로부터 제한된 도메인으로 자동번역 시스템을 특화하는 절차가 소개된 바 있다[5]. 이 도메인 특화 방법을 웹에서 특허로 영한 번역기를 특화할 때 사용하였는데 다음과 같은 6단계로 구성되었었다:

- 웹에서 특허로 특화한 방법(이후 '초기 도메인 특화 방법'으로 명명함)
  1단계: 해당 도메인으로부터 대량의 문서를 수집.
  2단계: 수집된 도메인 문장들을 대상으로 해당 도메인의 언어학적 특성을 분석.
  3단계: 도메인 문서로부터 미등록어 후보를 자동으로 추출하여 대역어를 반자동으로 구축.
  4단계: 도메인 특화된 대역 패턴을 수동으로 구축.
  5단계: 번역엔진 모듈을 특화.
  6단계: 전문 번역가에 의한 수동 평가

특허문서를 대상으로 초기 도메인 특화 방법이 적용되기 전과 후의 개선 결과를 비교하여 보았다. 특허 문장 중에서 200 문장을 임의로 추출하여 웹문서 영한 번역기를 적용한 결과와 특허문서 영한 번역기를 적용한 결과를 비교하여 보았다.

표 1. 초기 도메인 특화 방법 적용 전과 후

단계	항목	적용전	적용후	비고
3단계	신규 용어 구축	836,000	2,052,604	13개월, 증가량:1,216,604
4단계	신규 패턴 구축	39,127	50,124	18개월, 증가량: 11,087
5단계	태깅 정확률	95.85%	99.62%	
	파싱 정확률	69.00%	85.00%	
	대역어 선택 정확률	71.70%	92.40%	
6단계	번역률(수동평가)	54.25%	82.20%	

3. 초기 도메인 특화 방법의 문제점

초기 도메인 특화 방법에 의해 구축된 영한 특허 번역기를 가지고 특허 문서가 아닌 기술 논문을 평가하여 보았다. 기술논문 문장 중에서 200문장을 임의로 선정하여 평가한 결과는 다음과 같았다.

표 2. 특허 번역기에 의한 기술논문 자동번역 결과 오류 분석

항목		오류수	%
번역엔진	태깅	10	6.10%
	파싱	28	17.07%
	대역어 선택	3	1.83%
	생성	15	9.15%
번역지식	사전	77	46.95%
	패턴	23	14.02%
기타		8	4.88%
전체		164	100.00%
번역률(수동평가)			78.63%

표 2의 오류 분석을 통해 알 수 있는 것은 첫째, 사전과 패턴 오류를 합하면 60.97%로 기존 번역 지식 튜닝이 무엇보다 시급하다는 것이다. 두 번째로, 번역 엔진에 대한 튜닝의 자동화 방법이 필요하다는 것이며 끝으로 번역률 측정을 위한 비용과 시간을 절약하기 위해 자동평가 방법이 필요하다는 것이다.

4. 개선된 도메인 특화 방법: 특허 번역기를 기술논문 번역기로 특화하기

초기 도메인 특화 방법을 개선하기 위해 신규 도메인에 대해 기구축 용어를 반자동으로 튜닝하는 번역 사전 튜닝[6] 방법과 신규 도메인용 번역 패턴을 반자동으로 추가하는 코퍼스 기반의 패턴 확장[7] 방법과 번역엔진 튜닝 자동화를 위한 정답셋 기반 도메인 특화 방법이 추가되었다. 그 외 전문 번역가에 의한 수동평가 방법 외에 BLEU[8]라는 번역률 자동 측정 방법을 추가하였다. 개선된 도메인 특화 방법을 그림으로 보이면 다음과 같다 (상세한 내용은 [2]를 참조하기 바람):

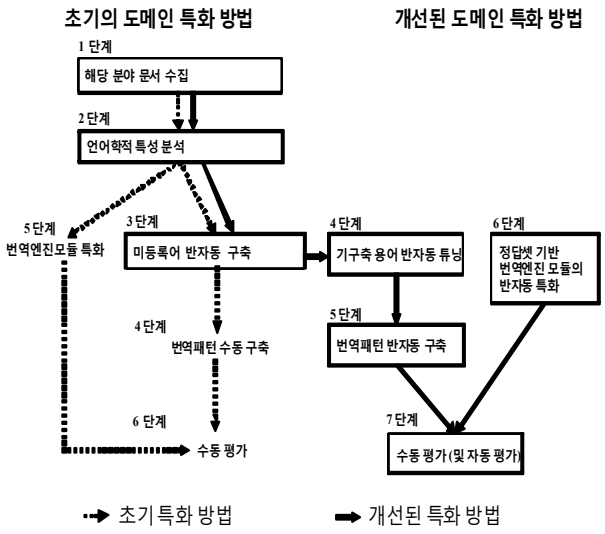

그림 1. 개선된 도메인 특화 방법

개선된 도메인 특화 방법은 총 7단계로 구성된다.

- 특허에서 기술논문으로 특화한 방법(이후 '개선된 도메인 특화 방법'으로 명명함)
  1단계: 해당 도메인으로부터 대량의 문서를 수집.

2단계: 수집된 도메인 문장들을 대상으로 해당 도메인의 언어학적 특성을 분석.
3단계: 도메인 문서로부터 미등록어 후보를 자동으로 추출하여 대역어를 반자동으로 구축.
4단계: 구축되어 있는 용어들을 도메인 문서에 나타나는 고빈도순으로 배열하고 병렬문장을 대상으로 대역어 반자동 튜닝을 실시.
5단계: 도메인 문서로부터 원문 패턴 후보를 자동으로 추출하여 대역 패턴을 수동으로 구축.
6단계: 정답셋에 기반하여 번역엔진 모듈을 반자동으로 특화.
7단계: 전문 번역가에 의한 수동 평가 뿐만 아니라 정답셋에 기반한 자동 평가를 실시.

초기 도메인 특화 방법에 비해 개선된 도메인 특화 방법의 개선 사항을 살펴보기 위해 기술논문 문장에서 400문장을 임의로 추출하여 수동번역률을 측정하였으며, 자동번역률은 문장당 5개의 참조문(Reference)을 부착한 1,000문장에 대해 BLEU 평가가 실시되었다. 각 단계별로 개선된 사항들은 다음의 도표와 같다.

표 3. 개선된 도메인 특화 방법 적용 전과 후

단계	항목	적용전	적용후	비고
3단계	신규 용어 구축	2,052,604	2,510,496	5개월, 증가량:457,892
5단계	신규 패턴 구축	50,124	74,337	6개월, 증가량:24,123
6단계	태깅 정확률	99.20%	99.27%	
	파싱 정확률	72.00%	82.00%	
	대역어 선택 정확률	79.00%	87.75%	
7단계	번역률(수동평가)	77.25%	81.10%	
	번역률(자동평가)	0.4946	0.5185	BLEU

5. 개선된 도메인 특화 방법의 타당성 실험: 전자우편 번역기를 메신저 번역기로 특화하기

개선된 도메인 특화 방법이 새로운 도메인에 대해서도 타당한 지를 실험하기 위해 기존에 개발되어 있는 전자우편 번역기를 Microsoft 메신저용 번역기로 특화하여 보았다. 개선된 도메인 특화 방법에 따라 메신저 대화체 문장의 언어적 특성을 분석하고 이 언어적 특성 분석에 따라 각 모듈을 특화하기 위한 항목들을 정리하면 다음과 같았다:

표 4. 영어 대화체 메신저 문장의 특화 결과

분류	특화 방법	예(특화전->특화후)
어투	번역 패턴 반자동 구축	Any girls wanna chat?:어떤 여자가 채팅하기를 원합니까?-> 채팅할 여자 누구 없어?
생략	번역엔진모듈 반자동 특화 -고빈도 생략구조 복원	You trying to go home -> You are trying to go home.
슬랭	미등록어 반자동 구축 -슬랭어 사전 구축 -축약 처리	lol->하하하 wb->돌아온 걸 환영해
의성어	번역엔진모듈 반자동 특화 -철자 오류 처리	awwwwwww->aw
대화체 대역어	기구축 용어 반자동 튜닝 -대화체 표현용 사전 대역어 수정	crazy:미친->열광하는
기호	번역엔진모듈 반자동 특화 -기호 및 이모티콘 처리	!!!!!!!!!!!!!!:!!!!!!!!!!!!!!!->!
철자오류	번역엔진모듈 반자동 특화 -철자 오류 처리	didnt->didn't
부르기	번역엔진모듈 반자동 특화 -철자 오류 처리	heyyyyy->hey
약어	미등록어 반자동 구축 -약어 사전 구축	NY->New York
고유명사	미등록어 반자동 구축 -고유명사 사전 구축	Buffalo Wings->버팔로윙
대문자	번역엔진모듈 반자동 특화 -소문자 전처리	WHY WOULD ANY ONE...->why would...

'분류'란은 메신저 대화체 문자의 언어적 특성을 항목별로 분류한 것이며, 특화방법은 각 모듈별로 튜닝된 항목을 말한다. 예는 이런 특화방법이 적용되기 전과 후의 결과를 나타낸다. 각 단계별로 개선된 사항들을 살펴보면 다음의 도표와 같다.

표 5. 개선된 도메인 특화 방법의 타당성

단계	항목	적용전	적용후	비고(모듈 특화 내용 포함)
3단계	신규 용어 구축	2,510,496	2,524,577	2개월, 증가량:14,081 (슬랭, 감탄사, 고유명사 등)
5단계	신규 패턴 구축	74,337	74,337	2개월, 증가량:1,661 (의미기반 번역패턴 위주)
6단계	태깅 정확률	-	99.10%	축약, 철자 오류, 이모티콘 처리
	파싱 정확률	-	74.00%	TM Fuzzy 매칭, 간투사, 생략 등
	대역어 선택 정확률	-	85.00%	'주어-동사','동사-목적어','형용사-명사' 연어DB 구축
7단계	번역률(수동평가)	80.4%	81.93%	
	번역률(자동평가)	-	-	

표 5에서 번역률(수동평가)은 영어권 Native speaker가 대화한 메신저 대화체 문장 중에서 임의로 자동 추출한 100문장을 대상으로 하였으며, 100문장의 평균 단어 수는 8.65 단어였다. 평가 방법은 5인의 번역가에게 0점(번역품 출력이 안됨)에서 4점(원어문의 의미가 그대로 전달됨)사이에 0.5점 대별로 평가 점수를 부여하게 하고 각 문장에서 최고/최저 점수를 제외한 3인 점수에 대한 평균으로 번역률을 계산하였다.

6. 결론

본 논문은 한국전자통신연구원 언어처리연구팀에서 2004년까지 개발하였던 웹문서 자동번역 시스템을 2006년부터 도메인별로 매년 상용화에 성공한 사례를 기술한 것이다. 본 논문에서는 상용화에 성공한 주요 요인인 도메인 특화 방법을 소개하였다. 2004년의 웹문서 영한 자동번역기를 2006년에 특허문서 영한 자동번역기로 도메인 특화할 때 사용한 특화 방법이 '초기 도메인 특화 방법'이라면, 2006년의 특허문서 영한 자동번역기를 2007년의 기술논문 영한 자동번역기로 도메인 특화할 때 사용한 특화 방법이 '개선된 도메인 특화 방법'이었다. '개선된 도메인 특화 방법'이 '초기 도메인 특화 방법'에 비해 우수한 점은 기구축 용어를 반자동으로 튜닝하는 번역 사전 튜닝방법, 신규 도메인의 번역 패턴을 반자동으로 추가하는 코퍼스 기반의 패턴 확장 방법, 번역엔진 튜닝의 자동화를 위해 정답셋에 기반한 도메인 특화 방법, BLEU에 의한 번역률 자동 측정 방법을 추가한 점이다.

아래의 그림들은 도메인 특화 방법에 의해 개발된 도메인별 영한 자동 번역 시스템의 스크린샷을 보여준다.

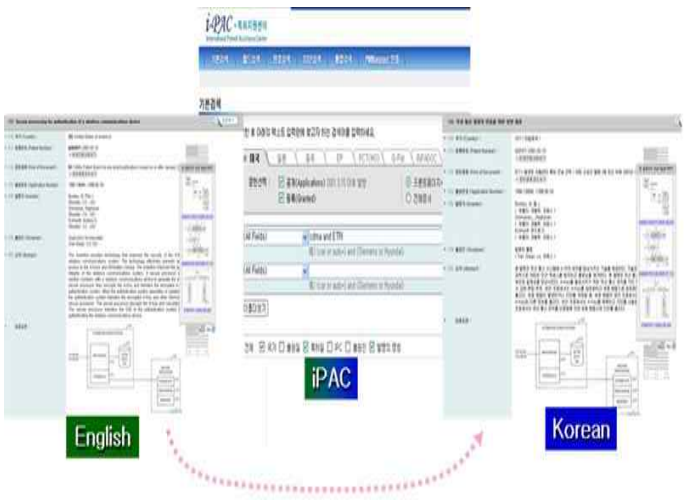

그림 2. 영한 특허문서 자동번역기

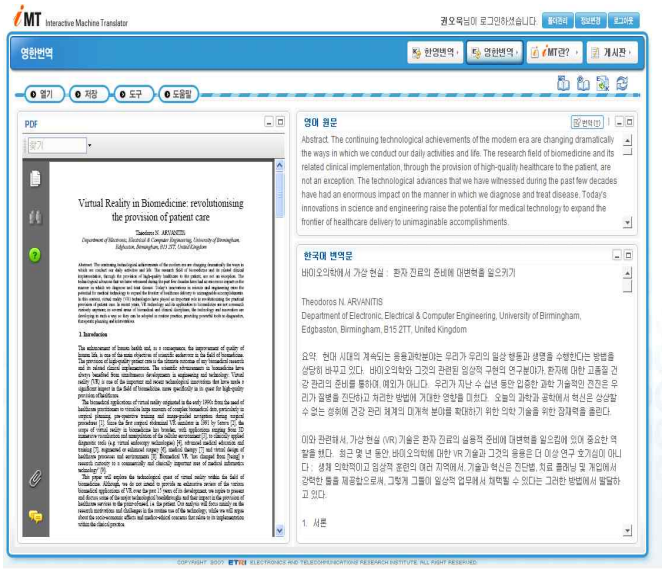

그림 3. 영한 기술논문 자동번역기

그림 4. 영한 대화체 메신저 자동번역기

참고문헌

[1] Sung-Kwon Choi, Oh-Woog Kwon, Ki-Young Lee, Yoon-Hyung Roh, and Young-Gil Kim. "Customizing an English-Korean Machine Translation System for Patent Translation", In Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation PACLIC 21), 2007, pp.105-114.

[2] Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh, Oh-Woog Kwon, and Young-Gil Kim. "How to Overcome the Domain Barriers in Pattern-Based Machine Translation System", In Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation(PACLIC 22), 2008, pp.161-168.

[3] Ki-Young Lee, Sung-Kwon Choi, Oh-Woog Kwon, Yoon-Hyung Roh, and Young-Gil Kim. "Domain Adaptation for English-Korean MT System: from Patent Domain to IT Web News Domain", In Proceedings of

the 22nd International Conference on the Computer Processing of Oriental Languages(ICCPOL 2009), 2009, pp.321-328.

[4] Oh-Woog Kwon, Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh, and Young-Gil Kim. "Customizing an English-Korean Machine Translation System for Patent/Technical Documents Translation", In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation(PACLIC 23), 2009, pp.718-725.

[5] Remi Zajac. "MT Customization". MT Summit IX Workshop, 2003.

[6] Ayan, N.F., B.J. Dorr and O. Kolak. “Domain Tuning of Bilingual Lexicons for MT”. CS-TR-4449, UMIACS-TR-2003-19, LAMP-TR-096, 2003.

[7] Yamada S., K. Imamura and K. Yamamoto. “Corpus-Assisted Expansion of Manual MT Knowledge”. In Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2002). 2002, pp.199-208.

[8] Papineni, K., S. Roukos, T. Ward and W.J. Zhu. “BLEU: a Method for Automatic Evaluation of Machine Translation”. In Proceedings of the 40thAnnual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, USA, 2002, pp.311-318.